1

A Strategy for Compression and Analysis of Massive Geophysical Data Sets

NASA launched its first Earth Observing System (EOS) satellite, Terra, into polar orbit on December 18,1999. Terra carries five instruments for studying Earth's climate systems over a six year period, and will produce vast quantities of data; more than much of the user community is equipped to handle. In geoscience, traditional strategies for coping with this problem are two-fold. The first is to work only with spatio-temporal subsets. The second is to work with low-resolution summaries typically created by partitioning data for a specified time period into 1° latitude by 1° longitude regions, and summarizing each region by its mean and standard deviation, or other simple statistics. The first strategy fails to take advantage of the global nature of Terra data, and requires researchers to know ahead of time where interesting phenomena exist. The second strategy fails to capture multivariate structure, and may aggregate away important high-resolution features.

This paper describes a method for summarizing these data in a way that approximately preserves high-resolution data structure while reducing data volume and maintaining global integrity of very large, remote sensing data sets. The method is under development for one of Terra's instruments, the Multi-angle Imaging SpectroRadiometer (MISR). The strategy is to partition data for each month into 1° latitude by 1° longitude spatial cells, and summarize each cell with a set of representative points and their associated frequencies. Each representative stands for some number of original observations, that number given by frequency. The combination of representatives and counts is a compressed version, or summary, of the original data. Researchers wishing to conduct global, exploratory analysis can do so using the compressed data with the understanding that results should be confirmed using appropriate portions of the original data.

The algorithm used to construct these summaries is a modification of the Entropy-constrained Vector Quantization algorithm (ECVQ) of Chou, Lookabaugh and Gray (1989), and is described in Section 1.2. First, however, Section 1.1 describes the MISR data stream. Finally, Section 1.3 provides an example analysis using compressed MISR data.

1.1 MISR Data

MISR (Diner, Beckert, Reilly, Bruegge et. al., 1998) is a set of nine cameras mounted underneath Terra looking down at Earth at nine view angles: 70.5°, 60.0°, 45.6°, and 26.1° aft; 0° (nadir), and 70.5°, 60.0°, 45.6°, and 26.1° forward along the spacecraft's north-south flight path. Each camera has four line arrays of 1504 pixel across the field of view and perpendicular to the flight track. The line arrays are each sensitive to one of four wavelengths: blue, green, red, and NIR (446, 558, 672 and 866 nanometers), and each pixel views a square region on the ground 275 meters on a side. Thus, one orbital swath on the daylight side of Earth tiles the view into disjoint, contiguous 275 meter spatial regions, and produces 36 radiance measurements for each one. Data for the nadir camera and the red bands in the other cameras are transmitted to Earth at full 275 meter resolution. Data for all other channels are averaged up to 1.1 km resolution on-board the spacecraft to limit data rate. The instrument does not take data as the satellite travels up the night side, so sequential orbits are separated. After 16 days 233 unique but overlapping orbits have completed covering the whole Earth. Every 234th orbit covers the same ground track as the first to within 20 kilometers.

MISR data processing takes this radiance data through several steps. First, data are geometrically and radiometrically calibrated to create the so-called Level 1B2 product. There is a seven minute lag between the forward and aft-most views of the same scene. Geometric rectification aligns the observations to produce 36 measurements (nine angles by four wavelengths) associated with the latitude and longitude of each pixel center. MISR produces about 40 GB per day of Level 1B2 data. Second, Level 2 data are created by converting these 36-vectors into geophysical quantities through complex science algorithms. For example, measurements taken within a 17.6 kilometer area are used to derive aerosol type and amount by matching observed radiances with those predicted by various physical models. Other quantities such as cloud height, wind direction and speed, and surface properties are derived at other spatial resolutions, typically 1.1, 2.2, and 35.2 kilometers. This second stage of processing reduces data volume by reducing spatial resolution, but increases data volume because many more than 36 geophysical variables are derived. MISR generates about 20 GB per day of Level 2 data.

The third processing step creates Level 3 monthly summaries by partitioning observations according to their membership in cells of a 1° latitude by 1° longitude spatial grid, and summarizing by grid cell. At the time of this writing the intention is to routinely produce compressed Level 3 data products derived from a select set of Level 2 geophysical variables. However, the method is also applied to portions of the Level 1B2 radiance data for research and analysis purposes. One such application is the topic of Section 1.3.

1.2 Monte Carlo Extended ECVQ

ECVQ is an iterative algorithm that groups data into a collection of disjoint clusters so as to minimize the loss function

$$L_\lambda = \sum_{n=1}^{N} \|y_n - \beta(\alpha(y_n))\|^2 + \lambda \left[-\log \frac{N(\alpha(y_n))}{N}\right], \qquad (1.1)$$

where y_n' is the nth row of an $N \times C$ data matrix representing one spatial cell in one month. $\alpha(y_n)$ is an integer indexing the cluster to which y_n is assigned, and $\beta(k)$ is the representative of the cluster indexed by k. $N(\alpha(y_n))$ is the number of data points (rows) assigned to y_n's cluster, and λ is a fixed constant. $-\log(N(k)/N)$ is positive and varies inversely with $N(k)$. Thus, even if $\|y_n - \beta(k_1)\|^2 > \|y_n - \beta(k_2)\|^2$, y_n could be assigned to cluster k_1 if the difference in the terms involving logarithms compensates. If $\lambda = 0$, L_λ is euclidian distance, and ECVQ is equivalent to the batch version of the K-means clustering procedure (MacQueen, 1967).

Briefly, the ECVQ algorithm works as follows:

1. Fix the maximum number of clusters allowed, K, and the compression parameter, λ.
2. Arbitrarily assign the y_n's to the K clusters by specifying initial values for $\alpha(y_n)$. Compute means and frequencies of these clusters, and denote them $\beta(k)$ and $N(k)$ respectively, for $k = 1, 2, \ldots, K$.
3. Reassign each y_n to the cluster with the smallest loss:
$$\alpha(y_n) = argmin_k \left\{ \|y_n - \beta(k)\|^2 + \lambda \left[-\log \frac{N(k)}{N}\right] \right\}.$$
4. Update $\beta(k)$ and $N(k)$ for all k.
5. Eliminate any clusters for which $N(k) = 0$.
6. Repeat steps (3), (4) and (5) until convergence.

The ECVQ solution has the property that the $\beta(k)$'s are the means of the y_n's they represent, a property call self-consistency by Tarpey and Flury (1996). However, assignment of data points to clusters is not nearest-neighbor in euclidian distance, and therefore does not minimize mean squared error,

$$\delta = \frac{1}{N} \sum_{n=1}^{N} \|y_n - \beta(\alpha(y_n))\|^2.$$

δ is also called distortion. The algorithm is guaranteed to converge in a finite number of steps, but not necessarily to either a local or global minimum. However, the solution does improve on the starting point providing a sensible summary of the y_n's in the sense described by MacQueen: "The

point of view taken in this application is *not* to find some unique, definitive grouping, but rather to simply aid the investigator in obtaining qualitative and quantitative understanding of large amounts of ... data by providing him with reasonably good similarity groups." (MacQueen, 1967, page 288.)

To apply ECVQ to large quantities of geophysical data, two modifications are made. First, since the algorithm is $O(n^2)$ and cell populations are large, a sample of M rows from each cell is chosen. ECVQ is applied to the sample, and an initial set of representatives, $\{\beta^*(k)\}_{k=1}^{K^*}$, obtained. This is the design step. Then each original data point in the cell is assigned to its nearest euclidian distance $\beta^*(k)$, empty clusters are deleted, and representatives and counts updated. This is the binning step. In other words, a preliminary set of representatives is determined from a sample, then the entire cell data set clustered using it. The ultimate set of clusters and counts thus reflects all the data, and is approximately nearest-neighbor. The resulting summary is denoted $\left\{\tilde{\beta}(k), \tilde{N}(k)\right\}_{k=1}^{\tilde{K}}$. This modification constitutes the Extended ECVQ (EECVQ) procedure.

The second modification addresses the fact that EECVQ is sample dependent, and $\left\{\tilde{\beta}(k), \tilde{N}(k)\right\}_{k=1}^{\tilde{K}}$ is subject to sampling variation. To account for this EECVQ is repeated S times using different random samples on each trial in the design step. This produces S summaries of the cell data, each one having a mean squared error δ_s. The best summary is the one having the smallest δ: $s_{opt} = argmin_s\{\delta_s\}_{s=1}^{S}$, and $\left\{\tilde{\beta}_{s_{opt}}(k), \tilde{N}_{s_{opt}}(k)\right\}_{k=1}^{\tilde{K}_{s_{opt}}}$ is selected to represent the original data. $\delta_{s_{opt}}$ is reported as a goodness of fit measure, and the entropy of the best summary,

$$h_{s_{opt}} = -\sum_{k=1}^{\tilde{K}} \frac{\tilde{N}_{opt}(k)}{N} \log \frac{\tilde{N}_{opt}(k)}{N},$$

is reported as a measure of descriptive complexity of the underlying data. Average mean squared error over trials, $\bar{\delta} = S^{-1}\sum_{s=1}^{S}\delta_s$, is also reported as an overall figure of merit. This procedure that embeds EECVQ in a Monte Carlo simulation is called Monte Carlo Extended ECVQ (MCEECVQ).

Finally, a value of λ must be selected. λ controls the level of compression. High values put a premium on the penalty $-\log N(k)/N$, cause summaries to collapse down to fewer, more highly concentrated clusters, and result in higher mean squared errors. Low values usually result greater numbers of clusters, higher entropies, and lower mean squared errors. Since entropy measures descriptive complexity of the output distribution, λ parameterizes the trade-off between distortion and complexity.

Choosing λ for any cell in isolation amounts to deciding how much compression one wants to achieve in that cell beyond that assured by fixing the initial number of clusters K, and how much mean squared error one

is willing to tolerate. When summarizing many cells in concert, it is desirable that distortions be as equal as possible across cells so differences in summaries reflect differences in data they represent, not differences how well summaries fit their parent data. As a consequence of the trade-off between mean squared error and entropy, this tends to produce summaries with entropies that reflect concentrations of mass in underlying empirical distributions. Figure 1.1 illustrates a simple example. The top two panels show two data sets drawn from mixtures of bivariate normal distributions. The middle panels show those data summarized using five clusters in each case: $K = 5$, $\lambda = 0$. Data from the top panels are shown on plot floors, the positions of the spike show locations of cluster representatives, and spike heights show cluster populations. Bottom panels show how these data are summarized by ECVQ with $K = 5$ and $\lambda = .04$. In the $\lambda = .04$ regime, high density regions are reflected by fewer, more massive clusters. The sums of squared distances between points and their nearest representatives are more nearly equal in the bottom panels than in the middle panels.

In practice one selects a value of K to limit the maximum size of the MCEECVQ output to $K \times B$, where B is the number of cells being summarized. This determines an overall level of mean squared error. Then, one selects λ to minimize the variance of $\bar{\delta}$'s across cells:

$$Var(\bar{\delta}) = \frac{1}{B} \sum_{b=1}^{B} (\bar{\delta}_b - \hat{\delta})^2,$$

where b indexes cell and $\hat{\delta} = B^{-1} \sum_{b=1}^{B} \bar{\delta}_b$. This requires testing various values of λ beforehand. If necessary because of data volume, this can be done using a subset of cells.

1.3 Application to MISR Data

To illustrate MCEECVQ, it is applied to MISR Level 1B2 data collected over the eastern United States on March 6, 2000. 15 of the 36 radiances are shown in Figure 1.2. The five panels shows the scene in three (red, green, blue) of the four spectral bands and from five of the nine view angles. All data used for this exercise have been average up to 1.1 km resolution. The data set partially depicted in Figure 1.2 has 491,044 observations, each representing a 1.1 km spatial region, and 38 columns: one for each view angle-spectral band combination, and latitude and longitude.

A frequent objective in the analysis of remote sensing data is to classify pixels in a scene. In Figure 1.2 water, ice (in the far eastern end of Lake Erie), clouds (over central New York state), vegetated land of various types and terrain, and haze are all evident. MISR's multi-angle observations provide a novel kind of information useful for these classifications. For example, haze is more obvious in the oblique angles because these views represent

longer paths through the atmosphere. The combination of 36 radiances is expected to provided better discriminatory power than single view-angle, multi-spectral data.

One way to classify the scene in Figure 1.2 would be to run a K-means cluster analysis on all 491,044 observations. Even choosing a modest value for K of, say ten, is computationally intensive and time consuming. Instead, the following procedure was used. First, the data were partitioned into 84 $1° \times 1°$ strata, and each strata summarized using MCEECVQ. MCEECVQ was applied with $K = 10$, $\lambda = 2$, sample size $M = 200$, and $S = 50$ trials. The samples used in the design step were first standardized using the grand means and variances for all 491,044 data points, and then projected into the space of the first ten principal components calculated from the grand correlation matrix. Ten principal components account for over 98 percent of the total variation in the data. MCEECVQ was applied to the transformed data, and representatives were re-transformed back to the original 36-dimensional data space before the binning step. $\lambda = 2$ was chosen after testing 15 values ($\lambda = 0, .5, 1, \ldots, 6.5, 7$) and determining $\lambda = 2$ minimized $Var(\tilde{\delta})$ across the 84 strata in the reduced space. This produced between one and ten representatives and associated counts for each spatial cell.

The MCEECVQ output contained a total of 479 representatives and counts, and 84 associated values of $\delta_{s_{opt}}$, $h_{s_{opt}}$ and $\bar{\delta}$. These are shown in Figure 1.3 along with the original cell populations, N, the numbers of representatives in the summaries, $\tilde{K}$. Note that cells with the largest values of N are not necessarily those with the largest numbers of representatives.

Second, a weighted K-means analysis with $K = 10$ was applied to the 479 representatives. Here, the 479 cluster representatives serve as data points to be clustered. Combining representatives and counts to form $\left\{\tilde{\beta}_{s_{opt}}(k), \tilde{N}_{s_{opt}}(k)\right\}_{k=1}^{479}$, an initial set of ten representative is selected at random with probabilities $\tilde{N}_{s_{opt}}(k) / \sum_{k=1}^{479} \tilde{N}_{s_{opt}}(k)$. These serve as the initial "supercluster" centroids, and each of the 479 cluster means is then assigned to the supercluster with the nearest centroid. Supercluster centroids are updated by computing the weighted averages of members with weights proportional to $\tilde{N}_{s_{opt}}(k)$. These steps are iterated until convergence. The weighted K-means procedure was repeated 50 times, and the solution with the smallest weighted mean squared error adopted. Weighted mean squared error is

$$\delta_{wtd} = \sum_{k=1}^{479} \left\| \tilde{\beta}_{s_{opt}}(k) - q(\tilde{\beta}_{s_{opt}}(k)) \right\|^2 \frac{\tilde{N}_{s_{opt}}(k)}{\sum_{k=1}^{479} \tilde{N}_{s_{opt}}(k)},$$

where $q(\tilde{\beta}_{s_{opt}}(k))$ is the representative of the cluster to which $\tilde{\beta}_{s_{opt}}(k)$ is assigned in this second-stage K-means analysis. The resulting $K = 10$ supercluster representatives are taken as the ten types used to classify the scene.

Finally, each of the original 491,044 data points is assigned an integer between one and ten indicating which of the ten superclusters it is nearest in euclidian distance. The map in Figure 1.4 shows the resulting classification of the scene. The classification identifies the band of haziness (in red) that is barely visible in the nadir image but more apparent in the 70° forward view in Figure 1.2. Figure 1.4 also distinguishes sun glint on water off the southeast coast and in Lake Ontario seen in the 45° forward view, ice at the extreme east end of Lake Erie, and ice and clouds over Lake Simcoe in the upper left corners of the images in Figure 1.2. The ability to differentiate between pixel types that are indistinguishable at a single nadir view highlights the principal behind multi-angle imaging. This K-means analysis is an example of a procedure scientists are interested in conducting on data of this type, but which may be impractical if not for a volume-reduced version of the data that approximately preserve high-dimensional relationships.

1.4 Summary and Conclusions

This paper describes a randomized version of the ECVQ algorithm for creating compressed versions of large geophysical data sets. The technique is especially well suited to remote sensing data such as that obtained from MISR, since they are naturally stratified by geographic location, and have strong high-dimensional structure. The technique is demonstrated by partitioning a test MISR data set according to membership in 1° latitude by 1° longitude spatial regions, and compressing data in each region. The compressed data are a set of representative vectors and associated counts which can be thought of as multivariate histograms with variable numbers of bins, and bins with sizes and shapes that adapt to the shape of the data in high-dimensional space. The algorithm is applied to all regions using common values of algorithm parameters K and λ. K specifies the maximum number of representatives and is set to limit the size of the output to no more than K times the number of spatial regions. This determines the overall level of error between the summaries and their parent data. λ sets the level of compression over and above that resulting from the choice of K, and is selected so that entropies of mass distributions in the compressed data reflects concentrations of mass in the original 1° latitude by 1° data sets. This choice of λ also equilibrates the mean squared errors between compressed and original data across spatial regions, and therefore yields summaries that are of approximately equal quality. Compressed data are then used in place of original data in a cluster analysis to create a thematic map of the Appalachian region of the US as seen by MISR.

This exercise was performed on a relatively small amount of test data. Samples were used in the design step, but the full test data set is used for

binning in each trial of the simulation. This requires $S+2$ passes through all the data: one to collect samples, one for each trial, and one to finally bin the data. In larger applications, it may not be possible to scan the full data set multiple times. Another version of the algorithm is under development to address this problem. Also, choosing λ requires testing multiple candidate values, 15 in this case. This necessitates running MCEECVQ 15 times on each 1° latitude by 1° data set. Again, the strategy may not be practical for larger applications. An alternative is to test for the best λ on a subset of 1° latitude by 1° regions. For example, in a global application those regions indexed by latitudes and longitudes evenly divisible by five or ten degrees could be used.

The method described here is on its way to becoming an operational algorithm for compressing a portion of MISR Level 2 geophysical data products on a monthly basis. To a large extent, its efficiency and effectiveness will depend on the data themselves. How well MCEECVQ scales up, and what further modifications are necessary remain to be seen.

References

[1] Chou, P.A., Lookabaugh, T., and Gray, R.M. (1989), "Entropy-constrained Vector Quantization," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**, 31-42.

[2] Diner, David J., Beckert, Jewel C., Reilly, Terrance H., Bruegge, Carol J., Conel, James E., Kahn, Ralph H., Martonchik, John V., Ackerman, Thomas P., Davies, Roger, Gerstl, Siegfried A. W., Gordon, Howard R., Muller, Jan-Peter, Myeni, Ranga B., Sellers, Piers J., Pinty, Bernard, and Verstrate, Michel M. (1998), "Multi-angle Imaging SpectroRadiometer (MISR) Instrument Description and Experiment Overview," *IEEE Transactions on Geoscience and Remote Sensing*, **36**, 4, 1072-1087.

[3] MacQueen, James B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-296.

[4] Tarpey, Thaddeus and Flury, Bernard (1996), "Self-Consistency: A Fundamental Concept in Statistics," *Statistical Science*, **11**, 3, 229-243.

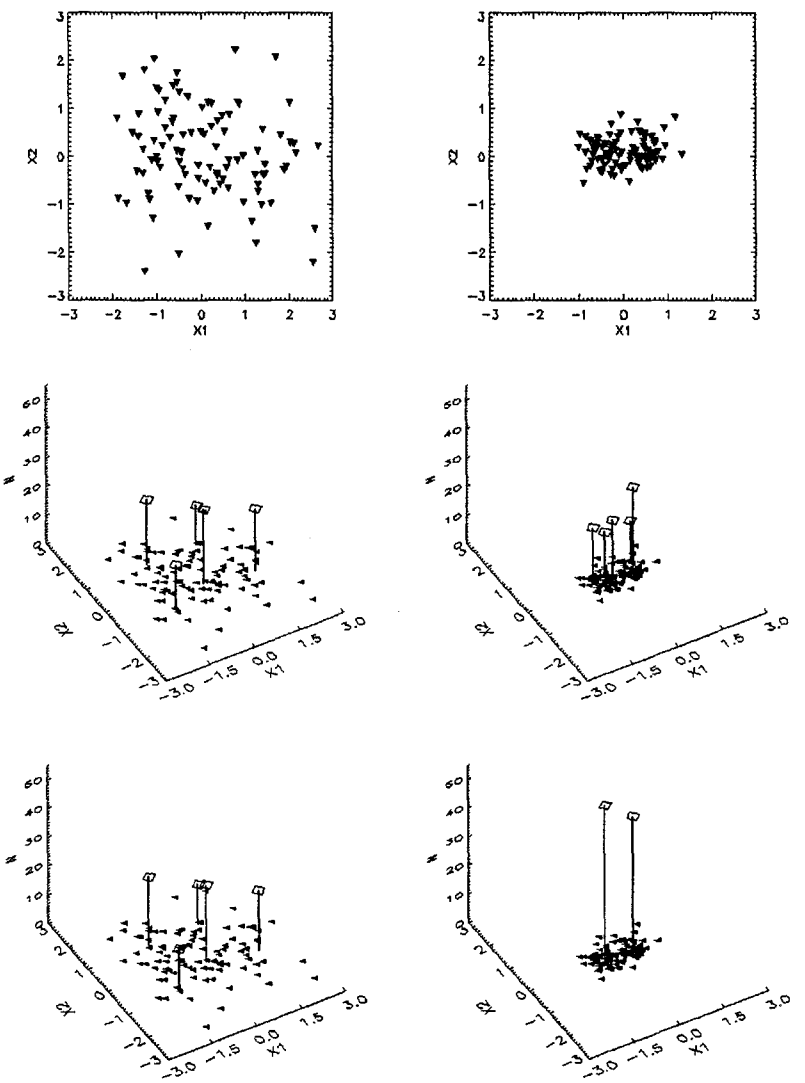

Figure 1.1. Top left: 100 observations from $f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$ where f_1 and f_2 are uncorrelated bivariate normals with $\mu_1 = (-.5, 0)$, $\mu_2 = (+.5, 0)$ and $\sigma_1 = \sigma_2 = 1$, and $\pi_1 = \pi_2 = .5$. Top right: same as top left except $\sigma_1 = \sigma_2 = .3$. Middle left and right: summaries of data with $K = 5$ and $\lambda = 0$. Bottom left and right: summaries of data with $K = 5$ and $\lambda = .04$.

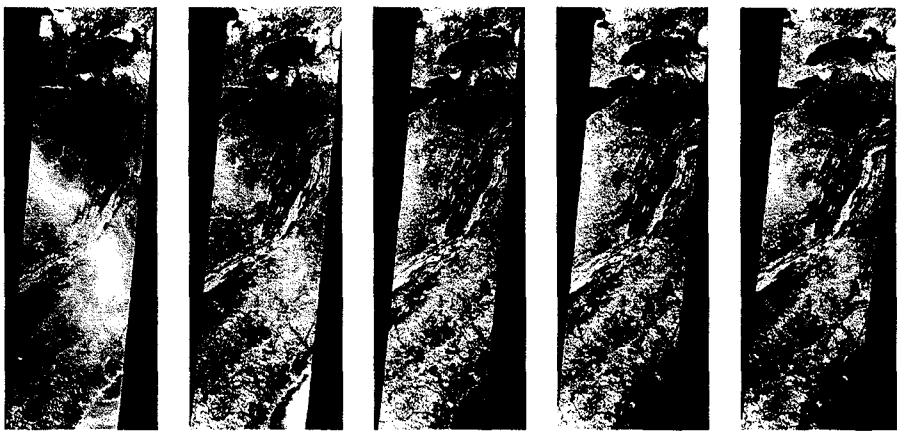

Figure 1.2. Left to right: 70.5° forward, 45.6° forward, nadir, 45.6° aft, and 70.5° aft.

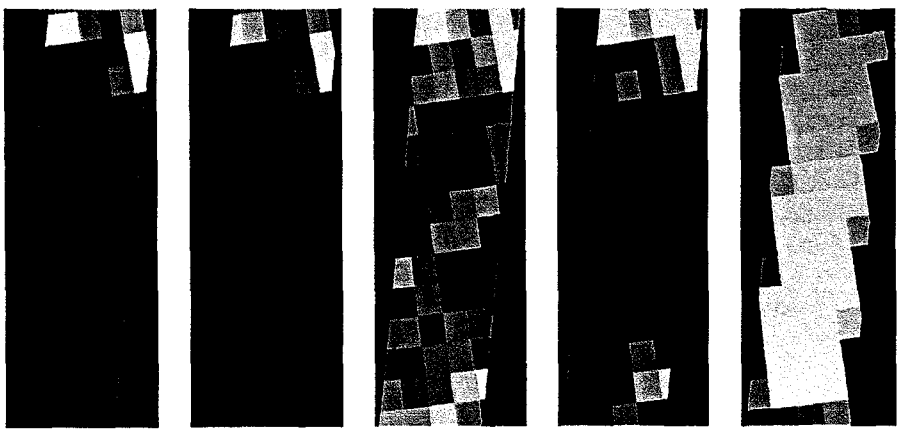

Figure 1.3. Left to right: Average cell mean squared error (over trials) as a proportion of average cell squared data norm; best cell summary mean squared error as a proportion of average cell squared data norm; best cell summary entropy; best cell summary number of clusters; cell population.

Figure 1.4. MISR thematic map of the Appalachian scene.